

TweepFake: About detecting deepfake tweets

Year: 2020 | Citations: 255 | Authors: T. Fagni, F. Falchi, Margherita Gambini, Antonio Martella, Maurizio Tesconi

Abstract

The recent advances in language modeling significantly improved the generative capabilities of deep neural models: in 2019 OpenAI released GPT-2, a pre-trained language model that can autonomously generate coherent, non-trivial and human-like text samples. Since then, ever more powerful text generative models have been developed. Adversaries can exploit these tremendous generative capabilities to enhance social bots that will have the ability to write plausible deepfake messages, hoping to contaminate public debate. To prevent this, it is crucial to develop deepfake social media messages detection systems. However, to the best of our knowledge no one has ever addressed the detection of machine-generated texts on social networks like Twitter or Facebook. With the aim of helping the research in this detection field, we collected the first dataset of real deepfake tweets, TweepFake. It is real in the sense that each deepfake tweet was actually posted on Twitter. We collected tweets from a total of 23 bots, imitating 17 human accounts. The bots are based on various generation techniques, i.e., Markov Chains, RNN, RNN+Markov, LSTM, GPT-2. We also randomly selected tweets from the humans imitated by the bots to have an overall balanced dataset of 25,572 tweets (half human and half bots generated). The dataset is publicly available on Kaggle. Lastly, we evaluated 13 deepfake text detection methods (based on various state-of-the-art approaches) to both demonstrate the challenges that Tweepfake poses and create a solid baseline of detection techniques. We hope that TweepFake can offer the opportunity to tackle the deepfake detection on social media messages as well.